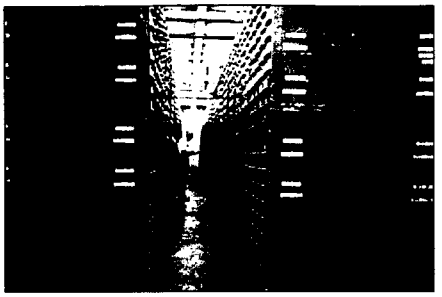
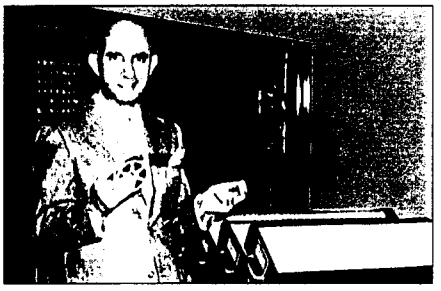
1940's

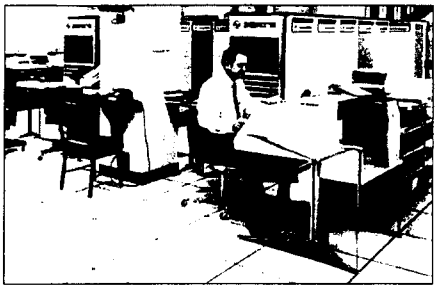


1950's

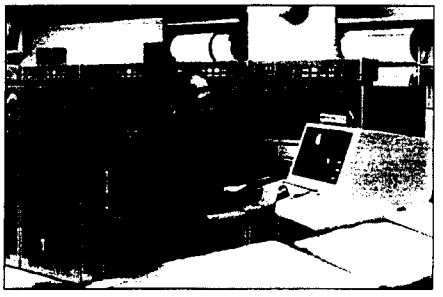
1960's



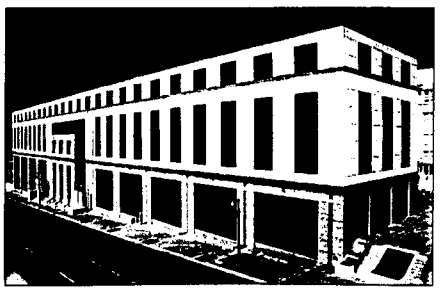
1970's



1980's



1990's



DATA RESCUE AT THE NATIONAL CLIMATIC DATA CENTER PROGRESS AND CHALLENGES

August 1995

NOAA

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

U.S. DEPARTMENT OF COMMERCE

Data Rescue at the National Climatic Data Center

Progress and Challenges

August 1995

Peter M. Steurer

On the Cover

The pictures on the front cover represent the evolution of data management activities at NCDC over the last 50 years. In the 1940s and 1950s, NCDC was involved with archiving massive quantities of climate data on several hundred million punchcards. The first picture shows a portion of the punchcard library. The second picture shows the "weather girls" digitizing paper records onto punchcards. During the 1950's almost 75 keypunchers labored every day to keep up with the inflow of data. The 1960s saw a need to preserve critical storage space and also rescue the aging punchcards. Helmut Landsberg, Weather Bureau Director of Climatology, is show in the third picture displaying the film output from one of NCDCs first data rescue activities . Landsberg is holding FOSDIC film and the machine behind him was called FOSDIC 1. This machine transferred punchcards to 16mm film resulting in storage space savings of 150 to 1. A different machine transferred the images from film back to punchcards, and in the 1970s, from film to magnetic tape. The next two pictures are mainframe computers used at NCDC during the 1970s and 1980s, respectively. The first mainframe is an RCA Spectra and was used in the rescue/migration from punchcards and FOSDIC film to 9 track magnetic tape. The second mainframe is a UNIVAC 1100 which replaced the Spectra in 1979. A UNISYS mainframe is still used at NCDC today primarily for data management and the rescue/migration of 9 track tapes to cartridges. For the past five years, NCDC has been transitioning to an Open system environment for all computer systems. This transition should be completed by 1998. The final picture is Asheville's new Federal Climate Complex building completed in 1995. This new facility has state-of-art environmental controls designed for the archival of data. The movement of NCDC records into the new facility will greatly expand the useful life of the archives but will not alleviated the deterioration of the collection that had already occurred.

TABLE OF CONTENTS

List of Figures

List of Tables

EXECUTIVE SUMMARY

This report is a comprehensive analysis of environmental data rescue at the National Climatic Data Center. Over the last 5 years, data rescue activities have resulted in a significant reversal in the decline of NCDC's data archives. However, 26 rescue tasks still remain to be completed due to the large volume and diversity of NCDC data.

Data rescue is defined as a preservation process where important old and new data are safeguarded for the present and perpetuated through the future. Three broad rescue categories at NCDC are (1) digital data with sub-categories of aging magnetic media, FOSDIC, and new data streams; (2) non-digital data with sub-categories of paper and film records; and (3) metadata with sub-categories of station history and data set documentation.

NCDC is the largest and most diverse environmental data center in the world containing over 90 percent of NOAAs data. The NCDC data resource is a cornerstone for the prediction of future events which affect the world's environment and economy. Decision makers use these data to solve problems ranging from assessing world food supplies to climate change. Billions of dollars have been spent to collect and process these irreplaceable data. The collection needs to be managed and preserved as a natural resource.

We propose to rescue only those data that have continuing value. This process of prioritization identifies unique data; establishes scientific requirement; identifies mission versus non-mission related data; considers the physical condition of the media; and evaluates the overall cost. To ensure that the data rescue process remains a dynamic facet of data management, quality assurance during the rescue process and the performance of periodic reassessments are proposed. A random sampling mechanism is recommended along with the development of a Comprehensive Records Control Schedule.

A total of 26 separate rescue tasks are identified at NCDC. To accomplish these tasks, the incremental cost over and above NCDC base funding is $2.7 million for five years. It is recommended that a rescue administrator position be added at NCDC to manage these tasks and associated contract work.

1.0 SCOPE: Definition of the Data Rescue Challenge at NCDC

Data rescue is a preservation process where data of continuing value are safeguarded for the present and perpetuated through the future. It is not a static one-time effort to resolve an urgent problem but a dynamic facet of data management that requires continuous planning and resources. Data rescue efforts involve not only the maintenance of existing data but also the ability to manage the processing of large volumes of data generated from new earth observing systems. Finely meshed within all aspects of data rescue is the need to prioritize to ensure that resources are used only on environmental data that have continuing value.

NCDC is the largest and most diverse environmental data center in the world. Data archives at NCDC contain a treasure trove of meteorological/climatological information. The collection represents this nation's heritage in that it contains the history of meteorological observations since the 1700s. The collection is so extensive that it has been estimated that almost every meteorological observation ever taken in the United States is available in some form in NCDC's archives. These data have proven to be extremely useful in predicting future events that affect the world's economy and environment. The data have become critical to the scientific community and policy makers in regard to global climate variability and trends.

Prior to the creation of the ESDIM Program in 1990, the large majority of this data legacy was at great risk. Budget limitations during the last decade and archive storage facilities with poor environmental controls placed these data in danger of being lost forever. New data streams originating from NWS modernization efforts exponentially magnified data preservation efforts due to the overwhelming amount of new data entering the process.

ESDIM has resulted in a significant reversal in the decay of environmental data. In addition, NCDC moved to a new building in 1995 with environmental controls designed for data archiving. However, due to the volume and diversity of NCDC data, much rescue work remains to be done.

Environmental data at NCDC can be grouped into three broad rescue categories. These are 1) digital data with sub-categories of aging magnetic media, new data streams, and FOSDIC; 2) non-digital data with sub-categories of paper and film records; and 3) metadata with sub-categories of station history and data set documentation. Figure 1 depicts examples of some of the older media used at NCDC that are in need of rescue.

Figure 1: *Archive media at NCDC requiring data rescue. Clockwise from the top: 9 track round tape, 16mm FOSDIC, 35mm film, and an 1870's logbook.*

1.1 Digital Data

NCDC's digital holdings currently comprise over 450,000 magnetic media units containing over 220 terabytes of information. It is estimated NCDC archives contain over 90 percent of all of NOAA's data. These data have been accumulated over the last 50 years. The largest increase has occurred since the mid-1970s with the advent of digital satellite data. Figure 2 represents the growth of NCDC's digital data base.

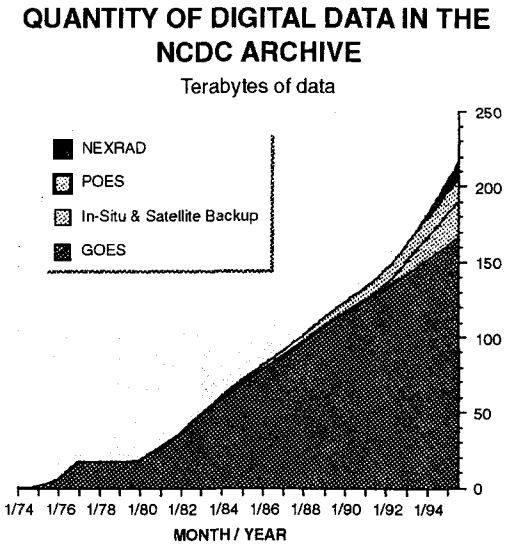


Figure 2: *Growth of NCDC digital data. Prior to 1974, NCDC archives contained less than 50 gigabytes of data consisting of climate in-situ data stored on several hundred million punchcards.*

In various studies in the 1980's and 1990s, GAO, IG, and NARA found that many of the digital media on which data were stored had exceeded their life expectancy (some media were over 25 years old) and that the NCDC was not performing sampling of digital media for integrity and stability in an attempt to avoid data loss. In addition, all satellite data existed only on single digital media without off-site back-up copies as required by NARA for security. As a result of ESDIM, random sampling is being performed, over 222,000 aging tapes have been migrated to stable media, and corresponding off-site back-up copies have been made for all but the GOES archive. However, now that most of these data have been rescued they must be maintained by periodically migrating to newer and more efficient media. This is a continuing responsibility of data management which prevents a rescue crisis from occurring again.

New data streams are resulting in a data explosion. Data from these new earth observing system are generated operationally. They cannot be managed with irregular funding resources but need to be included as part of the system design planning process.

Data streams from the National Weather Service are expected to bring over 100 terabytes of data per year to NCDC by 1996. The NEXRAD and ASOS modernization efforts started in the 1980s with the gradual commissioning of stations throughout the United States beginning in the early 1990s. NEXRAD involves the deployment of WSR-88D Doppler radars and ASOS consists of automated observations of surface weather conditions that were previously recorded manually.

Other new data streams which supplement the ASOS observations include lightning data from a nationally operated network and GOES cloud cover data to supply cloud information above 12,000 feet. Looming on the horizon is the EOSDIS archive which will have data volumes on the order of petabytes and has the potential of being managed in some form by NOAA.

An older digital media still available at NCDC is the FOSDIC collection. FOSDIC is 16-mm film that was used in the 1960s as a method to preserve deteriorating punchcards on microfilm. Physical storage space of cards was a major concern in the early days of automated data management. The FOSDIC migration process reduced the file space required to store the punchcards by approximately 150 to 1. Punchcards could be recreated from the film when needed. Later FOSDIC machines in the 1970s allowed the film to be converted directly to magnetic tape. In the 1970s' a project to convert the film to magnetic tape was begun. Before the effort could be completed, the

support for the FOSDIC system was abandoned leaving 64 million records on film and unreadable. Under the ESDIM program, NCDC has initiated a rescue process to transfer the FOSDIC film into standard digital format magnetic tape. A FOSDIC machine is shown in Figure 3.

1.2 Non-Digital Data

Non-digital data represents one of the most expensive data rescue challenges at NCDC due to the enormity of the collection and the urgency to save the rapidly deteriorating media. NCDC's non-digital holdings are divided into two categories: paper and film records. To a large extent, the same type of data are stored on these media types. The film represents a past migration effort that was performed to preserve the paper records. The film is now rapidly deteriorating and is in imminent danger of being lost forever. An

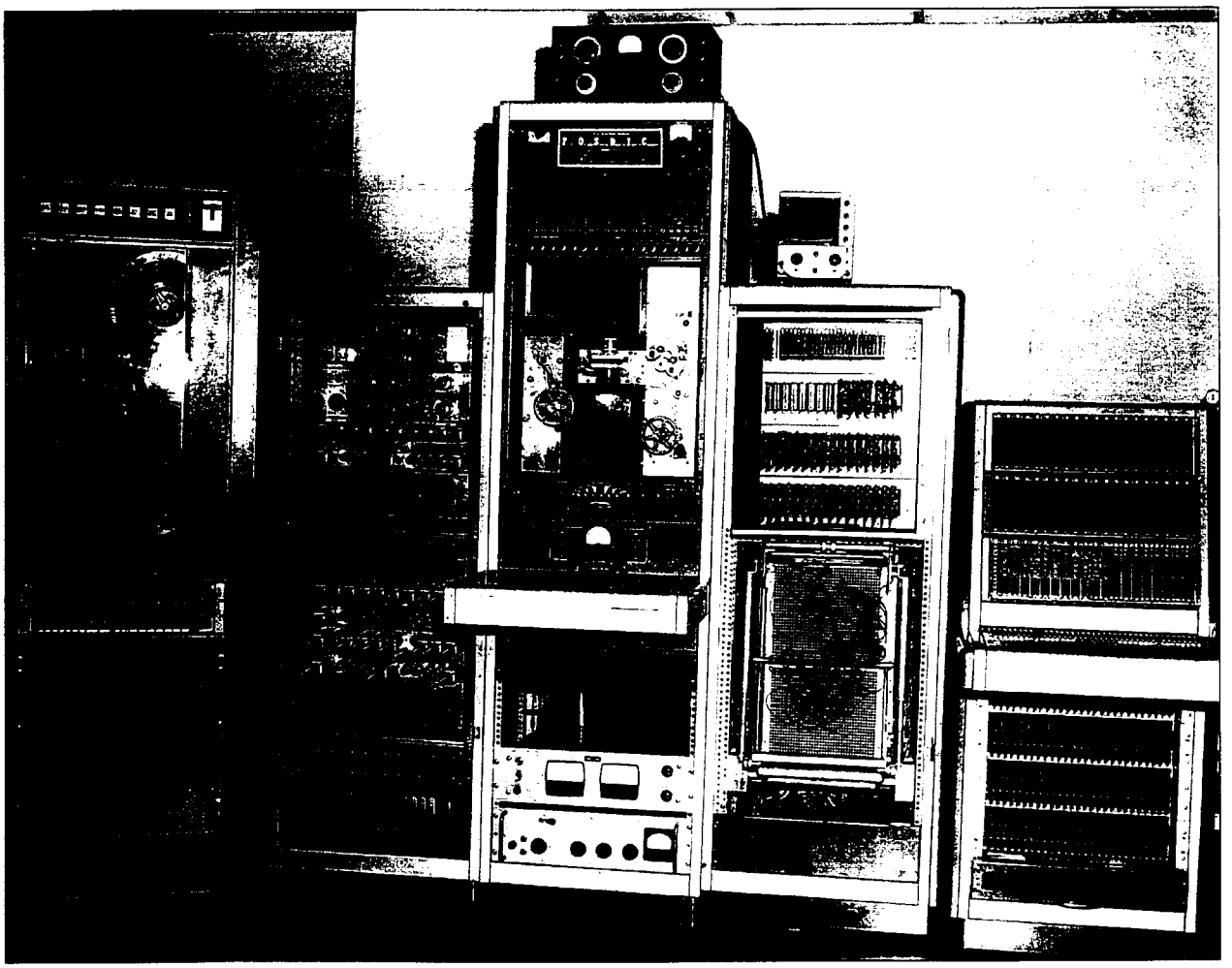


Figure 3: *FOSDIC 2 machine circa 1974. The center machine is reading punchcard images off 16mm FOSDIC film and the tape drive on the far left is recording these card images as 80 column digital records. Several hundred million punchcards, representing billions of dollars in keypunch labor, were processed through this machine during the 1970s. At the time, these pre-1970 data represented almost the entire climate in-situ data base for the United States. About 64 million cards remain to be done.*

example of this deterioration is shown in figure 4. Many of the paper records are also in danger due to deteriorating paper or fading ink.

The paper records consist of original manuscript forms and autographic charts recorded mainly over the last 100 years. These records consist of original observation forms which have one or multiple weather parameters recorded at some location on the earth for some time interval. The manuscript observations are usually manually observed and transcribed onto paper forms by humans. Autographic records are in the form of graphs/strip charts that are created by machines. It is estimated that NCDC's manuscript archive consists of 200 million paper records held in over 100,000 archive boxes which fill 50,000 linear feet of shelf space.

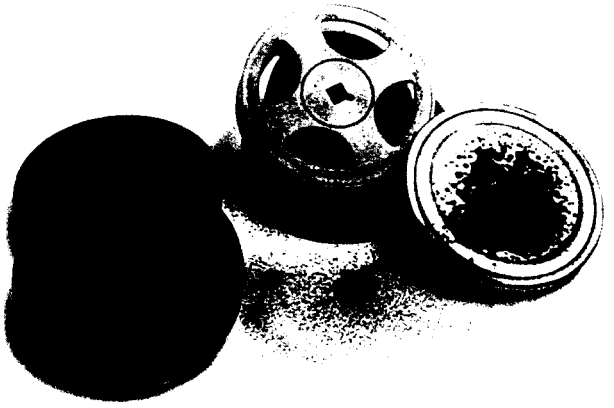


Figure 4: *Part of NCDC's Micrographics collection. The film is outgassing an acid that has caused the metal can to rust and the paper storage box (not shown) to disintegrate.*

Film or micrographic records consist of microfilm, 16mm film, and 35mm film. From the 1950s through November 1977, NCDC operationally filmed paper records using 16mm and 35mm acetate-based film. Many of the paper records were destroyed after filming. Since 1977, the media of choice has been microfilm. NCDC's film collection consists of over 2 million microfiche and 100,000 rolls of 35mm and 16mm film. The microfilm is on stable media but the roll film is rapidly degrading with less than a few years of useful life remaining.

The 1995 move of NCDC into a new facility brought the physical space where these records are archived into conformance with NARA standards. Also, a new archive/access system was developed around the same time which resulted in the first ever comprehensive inventory of NCDC's manuscript collection. In support of COADS, NCDC also rescued 3.5 million surface marine records by digitizing the information on manuscript forms.

1.3 Metadata

Metadata, or data about data, describe important aspects about a particular data set. In almost all instances, a data set has little value without accompanying metadata. NCDC metadata can be broadly grouped into the categories of data set documentation and station history.

Data set documentation contains descriptive information about a particular data set. This consists of such information as record format, data sort, sensors used, data quality, and known biases or problems. NCDC has several hundred data sets, all with some form of data set documentation. Unfortunately, complete documentation and descriptive information have been prepared only for the larger and more contemporary data sets. Documentation is incomplete or exist in single copy paper media for many of the older and potentially important collections. The documentation that does exist has been developed over many years and is in a multitude of formats making it difficult to use.

With support from the C&GC program, NCDC developed a PC based documentation builder that standardized the documentation process by presenting a series of 58 topics that would completely document a data set. All new documentation is placed in this standard format. The historical backlog remains to be completed.

Station history is associated with climate in-situ station networks. It consists of historical information related to a particular site such as the station name, latitude, longitude, elevation, instrumentation and exposure, observation schedules, and site characteristics. Changes to these site specific parameters can introduce inhomogeneities into the climatological data significant enough to cause erroneous conclusions in scientific work. NCDC has operationally maintained station history data in digital form since 1948 for stations that observe surface meteorological parameters. For the period 1890 to 1948, an ESDIM effort is resulting in the digitization of the manuscript station history records. Prior to 1890 the metadata still reside in manuscript form. Digital station history records are currently not comprehensive for other networks such as upper air and do not adequately address instrumentation and time of observation issues.

2.0 RATIONALE: The Importance of Data Rescue

Environmental data are the cornerstone for the prediction of future events which affect the United States' as well as the world's environment and economy. Accurate, accessible environmental data are critical to our understanding and description of global climate and regional and local environmental processes. These data form the basis for making decisions having economic and political consequences on local, regional, and global levels. The concern for climate change and its potential impacts are causing an increased need for NOAA data. Existing NCDC holdings of historical data, some dating back into the 1800's, as well as new data streams, represent the foundation needed to support climate applications and earth science studies.

Much of NOAA's environmental data represent a significant national and international resource that must be safeguarded and preserved. These data, which cost billions of dollars to collect, are distributed to and used by thousands of researchers in commerce, industry, science and engineering, national defense, and government agencies in applications ranging from estimating world food supplies to the understanding and the intelligent use of the environment. More importantly, these data are increasingly important to understanding the environmental changes on our planet. They assist scientists in answering questions on the extent and the potential impacts of global changes in the earth's environment, and on the well being and quality of life of future generations. These data are critical to research of ozone depletion, global climate warming, sea level change, drought, desertification, and the reduction in the diversity of living organisms. Much of these environmental data are irreplaceable; therefore, they are required to be managed and preserved as a valuable national resource.

3.0 METHODOLOGY: Generalized Data Rescue Plan

The rescue of environmental data requires a well-planned and organized process to ensure that data of continuing value are protected for current and future uses. At NCDC, a general approach has been used for all data at risk regardless of media type. As shown in section 4, the actual worksteps to accomplish data rescue for each media type can be vastly different but the generalized process flow outlined below is identical.

3.1 Data Rescue Prioritization

A common thread among all aspects of data rescue is the need to ascertain which environmental data have continuing value for current and future needs. The data rescue process then, requires a prioritization of data to ensure that resources are used only on the most critical data. The prioritization process can be divided into the five main areas shown in figure 5. These are a) the identification of unique data in the archive or planned for the archive, b) the scientific requirements for the data, c) the identification of mission

versus non-mission related data, d) the physical condition of the media upon which the data are stored, and e) the overall cost. Based upon an analysis of each of these prioritization areas, a determination is made for each data set or data base as to whether it should be rescued now, rescued in the future, or not rescued at all and possibly destroyed.

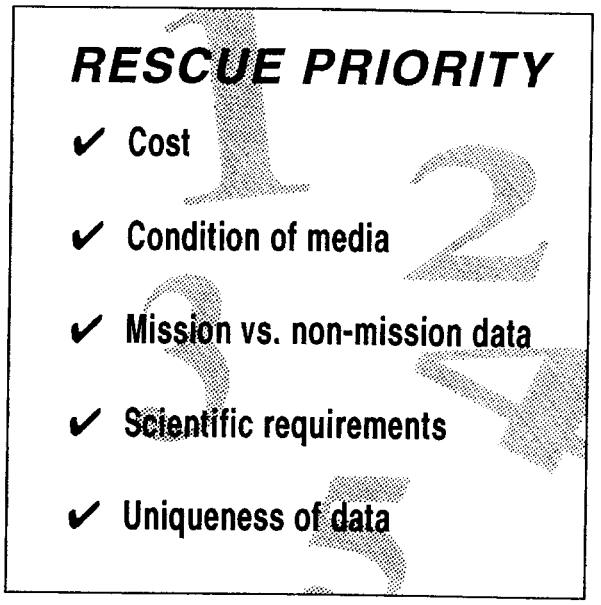


Figure 5: *Criteria for establishing data rescue priority. The cold facts of budgets have an overwhelming influence on any data rescue effort.*

The identification of unique data in the archive or planned for the archive appears to be a relatively simple task. However, there are many complexities. Duplicates and/or subsets of the data sets exist throughout NCDC archives. For example, the NCDC archive contains some data sets that are virtual duplicates of one another except that the format has been changed. Other data sets are subsets of data sets that have been produced for special projects. In some instances, the same data may be stored on different media types. These must be identified and only the unique sets should be rescued. An exception would be when the subset or odd format is still used and difficult or time consuming to reproduce.

The scientific requirement for the data is a broad issue that touches upon many areas. This requirement must not only address the current demand for data but anticipate the future need. It also addresses the requirement for high resolution data both temporally and spatially. Although data resolution is primarily a new data stream issue such as with NEXRAD and ASOS, it does exist with historical data sets such as satellite data. Historical significance is also an important issue. As an example, weather diaries from Thomas Jefferson or George Washington Carver may have some scientific value, but the historical value of these records is enormous. In determining scientific value, advice from data experts such as the NCDC Science Advisory Panel is solicited. Final determinations are made by the NCDC Director with input from the Data Administrator.

The identification of mission versus non-mission related data is a critical issue to establish during times of limited resources. A tremendous amount of scientific data currently exists, with exponential increases expected in the future. Since NCDC is currently asked to archive and distribute more data than it can possibly manage, data which is core to the organization's mission must be treated differently from non-mission related data. At NCDC, a determination is made by the Data Administrator as to whether or not a data set is core to the Center's functional responsibility. For larger volume data sets or data bases recommendations from NCDC's Science Advisory panel with concurrence from the NCDC Director is obtained.

Mission related data receive a much higher priority in rescue activities than non-mission related data. In addition, mission related data receive more extensive quality control and higher priority in conforming to NARA regulations on data management. These data sets are almost always archived on NARA approved media with primary and off-site back-up copies. Also, extensive metadata are maintained such as detailed data set documentation and station histories. Non-mission data may not be stored on NARA approved media and may not have

a back-up copy. In addition, the metadata and data quality assurance would be that provided by the data contributor with limited involvement from NCDC.

The physical condition of the media upon which the data are stored is a major consideration in determining data rescue priorities. If the media is stable with an extended useful life, data rescue can be postponed. An important exception is if the data have limited accessibility. In this case, migration to a newer media might solve both a future data rescue issue as well as a current data accessibility concern.

Unstable or deteriorating media requires immediate rescue action but may be overridden by other factors such as associated cost or low science value. At a very minimum, data on deteriorating media which have continuing value should be maintained in recommended environmentally controlled areas. This action may preserve or extend the life of the media allowing a future rescue action if resources should become available.

The overall cost of data rescue needs to be established to determine the economic feasibility of the rescue process. In fact, the cost of data rescue usually outweighs all the other factors described above in determining whether or not a data set is rescued. As an example, a data set with significant scientific value on a deteriorating media may not be rescued due to an exorbitant cost. The cold facts of budgets have an overwhelming influence on any data rescue effort.

3.2 Quality Assurance of Rescue Process

Quality assurance during the rescue or migration process is extremely important, and is the only means of ensuring that the process has been satisfactorily completed. This may be as simple as implementing the NARA sampling method for images on microfiche or as complex as a bit for bit digital data set comparison. But, it needs to be determined if the sampling method or the bit for bit comparison is adequate or complete enough. This in general is determined by the Data Administrator. However, quality control cannot be viewed as a one time function to be performed during the rescue or migration process only. A Data Center wide quality control activity must be an ongoing process.

3.3 Periodic Reassessments of Rescue Requirements

Data rescue is a dynamic facet of data management that requires continuous planning and resources. Periodic reassessments of rescue requirements are critical to ensure that data does not have to be rescued again in a crisis mode but migrated as part of an organized data management plan. This involves annual reassessments of the rescue process. and the development of a Comprehensive Records Control Schedule as shown in Figure 6.

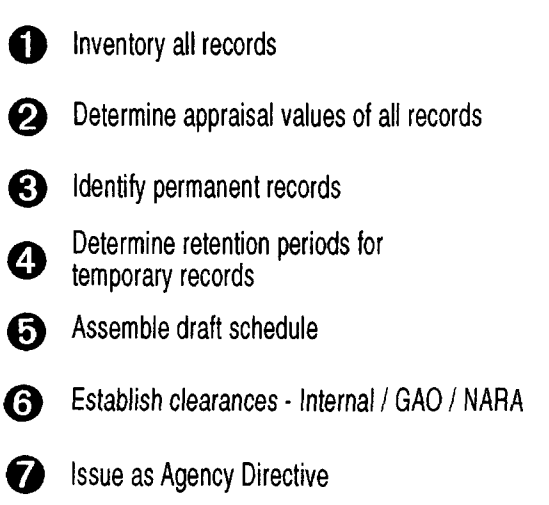


Figure 6: *Process flow of a Comprehensive Records Control Schedule; the lifeblood of organized data management.*

Annual reassessments are performed at NCDC for some data. As recommended by NARA, a three percent random sampling of digital data is performed annually. The purpose of a random sample is to determine patterns of data loss and develop correction

plans to prevent the loss. As an example, data loss may be occurring due to a periodic tape drive failure or a "bad" batch of tapes. Correction plans would involve repairing the tape drive or isolating the batch of tapes and recreating them from back-ups. A sampling procedure does not exist for NCDC's non-digital data collection and needs to be established to ensure that data does not deteriorate with age.

To prevent a future requirement of having to re-rescue data, a Comprehensive Records Control Schedule needs to be developed at NCDC. This is a formal document that describes all records of an agency, specifying those records to be preserved which having archival value. Once all records have been identified a comprehensive Disposition Program needs to be established which involves the development of standards, procedures, and techniques for managing the longevity of records. This would involve developing disposition schedules for all data sets which among other things, would determine retention schedules. These documents exist at NCDC but are not comprehensive or current.

4.0 THE DETAILS: Data Rescue Accomplishments and Challenges

The current situation at NCDC is both good and bad. The ESDIM program permitted tremendous breakthroughs in the area of data rescue. However, due to the diversity and large volume of data at NCDC much work remains to be done. Listed below in detail are the data rescue accomplishments that have been made over the last 5 years under the ESDIM program. Also listed below is the vital work that still needs to be performed. A total of 26 rescue tasks have been identified.

4.1 Digital Data

4.1.1 Magnetic Media

Accomplishments: NCDC operates as an Agency Records Center, so designated by GSA in 1951, then the parent organization of NARA. Therefore, the NCDC is obligated to maintain its records according to the guidelines established by NARA. Due to budget limitations during the last decade, NCDC has not been able to follow these guidelines in managing its data base. The GAO in its review of NCDC in 1989 (Report of November 1990) found that many of the digital media on which data were stored had exceeded their life expectancy (some media were over 25 years old) and that NCDC was not performing sampling of digital media for integrity and stability in an attempt to avoid data loss. In addition, all satellite data existed only on single digital media without off-site back-up copies as required by NARA for security.

In October 1990, NCDC initiated an effort under the ESDIM Program to rescue satellite and in-situ data at risk of being lost. The rescue activity involved migrating data from round 800, 1600, and 6250 BPI tapes to 3480 tape cartridges in order to save the data and also reduce critical storage space. Since the ESDIM program began in 1990, NCDC has rescued over 222,000 magnetic tapes. In addition, a corresponding back-up cartridge has been created and is now stored off-site. The rescue cost by reel has been between $20. and $25. per input magnetic tape which includes both a primary and back-up output 3480 cartridge copy. Table 1 depicts the data base categories at NCDC which have or will be rescued.

In 1991, NCDC also initiated a program to annually perform a 3% random sample of the digital data base to verify the integrity of the media as required by NARA. This activity will be continued as required by regulatory statutes. The cost of sampling continues to increase as a result of increasing data holdings from new data streams and the creation of back-up copies for security.

Vital Work: As shown in Table 1, a few data bases exist that still reside on round tapes and will be migrated in 1996. After 1996, magnetic media that were migrated in the early 1990's begin to exceed their expected life cycle according to NARA and come back into the inventory for migration

again to a newer media. NCDC plans to use 3590 cartridge technology which has a capacity of 10 gigabytes. These cartridges will be placed in a mass storage system called HDSS to allow on-line access. A major rescue challenge exists with the GOES archive. These data exist as single copy media. The EOSDIS Pathfinder efforts have produced some back-up copies for limited time periods. A plan needs to be developed to provide off-site back-ups for the entire GOES archive.

4.1.2 New Data Streams

Accomplishments: Data streams from new earth observing systems have the potential to completing overwhelm data management activities due to the extreme volume of new data entering the system. New data streams that are affecting NCDC are primarily associated with those generated by NWS modernization efforts. Figure 7 depicts an estimated growth rate of NCDC's digital archive through the year 2000. Looming on the horizon is NOAAs involvement with EOSDIS which has the potential of bringing many petabytes of information to NCDC. The future ramifications of the EOSDIS archive are not addressed in this report.

The ESDIM program has allowed NCDC to develop processing systems to manage data from two major NWS modernization efforts: NEXRAD and ASOS. The NEXRAD program is resulting in the deployment of up to 175 advanced WSR-88D pulsed Doppler radar systems throughout the United States and at selected overseas locations. The data generated include Level III products and the Level II high

Table 1. Data base migration schedule.
(thousands of tapes)

Data Type	*1991*	*1992*	*1993*	*1994*	*1995*	*1996*	*Status*
POES Level 1B (10/91 - current)	-	16.4	16.4	16.4	16.4	16.4	operational
POES Level 1B (08/90 - 10/91)	-	4.7	12.8	-	-	-	completed
POES Level 1B (1986 - 08/90)	-	-	2.3	20.5	20.5	7.0	-
POES Level 1B (pre-1986)	26.0	10.0	-	-	-	-	completed
CZCS	-	12.0	13.0	-	-	-	completed
ISCCP (Retro)	-	4.5	2.0	-	-	-	completed
SSM/I (Retro)	-	2.6	1.0	-	-	-	completed
HRPT (URI)	-	-	-	6.8	-	-	completed
VHRR, SIRS, DMSP	-	-	-	-	-	8.6	-
HRPT British Recovery	-	-	-	-	-	5.0	-
POES Level 1B (Jenne, 12/78 - 85)	-	-	-	-	-	4.5	-
GOES (Products)	-	-	-	-	-	17.0	-
In-Situ	-	7.0	7.0	4.0	-	-	completed
Yearly Totals: (excluding backups)	**26.0**	**57.2**	**54.5**	**47.7**	**36.9**	**58.5**	

Table 1: *Digital data bases rescued at NCDC from 1991-1996. Figure shows the number of input tapes migrated to 3480 cartridges. Over 222,000 magnetic tapes have been rescued so far. The rescue will be completed in 1996. However, once these data have been rescued, they must be maintained by continually migrating to newer and more efficient media.*

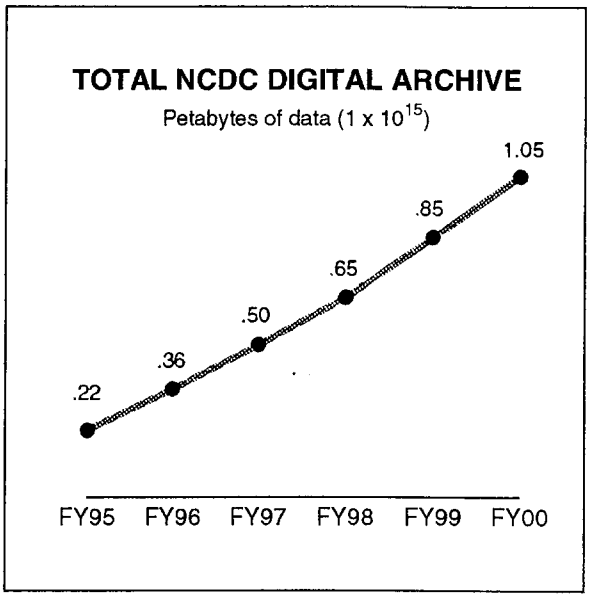


Figure 7: *NCDCs digital archive is expected to exceed 1 petabyte by the year 2000 due mainly to data from new observing systems. Not included in this estimate is NOAAs involvement with the EOSDIS archive which would bring many petabytes of data to NCDC.*

resolution. The Level II data consist of the three basic Doppler moments: reflectivity, radial velocity, and spectrum width. Data volumes from NEXRAD will approach 100 terabytes per year beginning in 1996.

As a result of ESDIM, NCDC has been able to procure several workstations and tape drives in order to process the massive NEXRAD data volume and produce back-up copies of all 8 mm tapes. An agreement was also reached with the University of Oklahoma to act as an off-site back-up for the NEXRAD archive.

ASOS data consist of automated observations of surface weather conditions that were previously recorded manually. The data include hourly observations and high resolution data. The high resolution data are 1 minute observations of weather parameters many of which were previously collected using chart recording instruments at manual observation sites.

While NEXRAD challenges are primarily concerned with data volume, ASOS challenges are associated with communications. In order to obtain high resolution 1-minute ASOS data, NCDC must communicate directly with the ASOS site as opposed to a centralized collection facility. NCDC has developed a distributed communications system which twice a day downloads hourly and high resolution data from each site. Currently, 100 sites are being accessed with future deployment of stations expected to be approximately 500.

Vital Work: Now that the systems to manage the processing of NEXRAD and ASOS data are in place, regular funding is required to continue data ingest operations. Ideally, this funding should be allocated during the planning process of any new data stream. But in reality, data management funds are usually the first to be cut during the budget process. It is vital that ASOS and NEXRAD continue to receive regular funding in future years to avoid data loss caused by the inability to process these data.

Additional tasks remain to be done and are outlined below:

1) Other new data streams exist which complement the ASOS observations and need to be preserved and made accessible. These are the lightning data and GOES cloud cover data.

Lightning data are collected from a proprietary national detection network system. The lightning data are essential to detect the occurrence of thunderstorms and maintain continuity with historical data since thunderstorm information is missing from ASOS observations. Unfortunately, when NWS awarded the contract for this system, there were no provisions for an archive of these data, thus risking their loss.

GOES cloud cover data are needed to complement the ASOS observation since ASOS ceilometers do not observe clouds above 12,000 feet. These data would also provide information which was unavailable in the past for surface observation sites. Cloud top information combined with cloud base could be used to determine thickness.

Also, more accurate cloud data could be obtained during periods of darkness or surface obstructed views.

2) Manually observed observations will soon become available which will supplement the ASOS observations. An ingest mechanism needs to implement to receive these data at NCDC in order to provide archival and accessibility.

4.1.3 FOSDIC

Accomplishments: Under the ESDIM program, NCDC has initiated a project to rescue vintage FOSDIC film. During the 1960's, in an effort to preserve deteriorating punch cards and save storage space, NCDC began a very futuristic program of placing the punch cards on 16mm microfilm using a FOSDIC system developed by the NBS. The FOSDIC process reduced the space required to store the records by a factor of 150 to 1.

Initially, images were recalled back onto to cards as needed. Later technology allowed the transfer of the microfilm images directly onto magnetic tape. In the 1970's, NCDC entered into a massive rescue project designed to recall all of these cards onto magnetic tape. Before the project could be completed, the NBS halted system support leaving 64 million cards on film and unreadable for the past 20 years.

In 1991 it was discovered that the "Father of FOSDIC" was still alive and under short term contract to the Bureau of Census as a consultant. A series of dialogues began with Census officials to explore the possibility of having them modify equipment and contract with NCDC to rescue the most vital of the records from FOSDIC film stored in vaults at the FRC in Atlanta.

Accomplishments under the ESDIM program have included investigative work to determine the feasibility of completing the task. After this was determined to be positive, funds were transferred to Census to allow for the development and equipment modification of the FOSDIC film reader. As of this time, approximately 100,000 images have been converted to magnetic media.

Vital Work: Now that the system is in place, the actual rescue production will increase significantly. It is expected that a total of 64 million images will be converted to magnetic media in 1996. Once the conversion is completed, the varied punch card formats that have been rescued to magnetic media will be segregated into data sets, documented, and made accessible in NCDC's magnetic media archive.

4.2 Non-Digital Data

4.2.1 Manuscript/Autographic Records

Accomplishments: Several GAO, IG, and NARA inspections have addressed the deteriorating conditions of NCDCs paper records and the need to improve security and environmental conditions. In 1995, NCDC moved into a newly constructed building that has brought the physical storage areas into conformance with established standards. However, the movement of the records into these new facilities has not alleviated the deterioration due to high usage, or that already under way because of the age of paper and fading of ink.

In 1995, a modern archive/retrieval system was implemented at NCDC which significantly improved access to the nation's historic climate records stored on paper media. The estimated 200 million paper records are held in over 100,000 archive boxes which fill 50,000 linear feet of shelf space.

As the world began its entry into the information age, the necessity for a modern archive/access system to hold NCDC's vast collection of data became increasing evident. Time-honored storage procedures, that worked well in the past, could no longer keep pace with the accelerating demand for rapid knowledge and access to NCDC's data. Automation was badly needed but could not be completed satisfactorily due to the enormity of the collection and inadequate resources.

The 1995 move of NCDC to a new Federal Climate Complex provided the spark that was needed to carry the management of NCDC's non-digital data

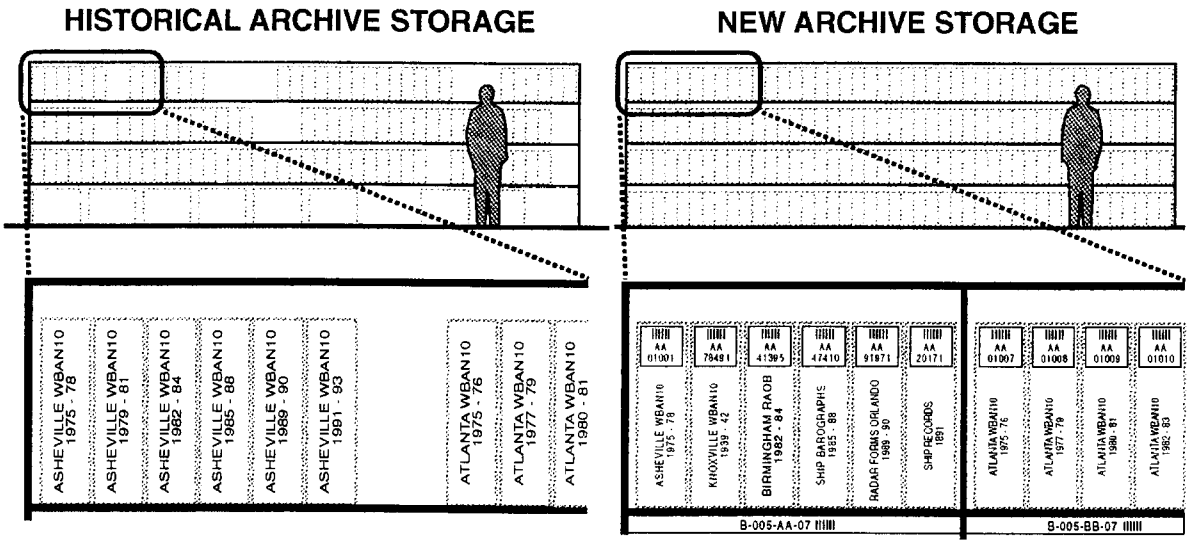


Figure 8: *Comparison of old and new manuscript storage systems. The left graphic represents archive storage used from the 1940's through 1994. The right graphic shows NCDC's new bar-coded archive/access system. The new system lessens the need for corporate knowledge in locating records. Inventory information is searchable via an on-line browser indicating the existence and location of a record box. The system allows for the random storage of boxes eliminating the labor-intensive need to shift boxes to maintain station/time sorts. The archive consists of 200 million records stored in 100,000 boxes.*

into the 21 Century. By combining the funds required to move the records to a new location with limited data management funds, an innovative approach was devised which resulted in a modern archive/access system.

As shown in figure 8, the new archive/access system uses bar-code technology to associate an individually bar-coded record box with a bar-coded shelf location. The inventory information identified with each record box is searchable via an on-line browser to indicate the existence of a particular data request and the shelf location. The system allows for the random storage of record boxes as opposed to the alphabetical time-ordered sort that was required in the past. The new system also eliminates the labor-intensive need for the constant shifting of record boxes which was required to maintain the alphabetical/time sort.

Another accomplishment is a highly detailed inventory of NCDC's Foreign Data Library. This library contains climate data publications from all countries in the world acquired operationally over the last 40 years. The collection has been cataloged to the publication level. Over 5,000 entries were placed in OCLC.

In support of COADS, NCDC also rescued 3.5 million surface marine records by digitizing the information on manuscript forms. These records are filling gaps in the historical records for the World War I and II years and are vital for climate change research. Additional records are planned for digitization in the future including the Maury collection

for the years prior to the 1860s.

Vital Work: The automated archive/ access system described above is a critical step in the rescue process: that of providing the first automated inventory of NCDC's on-site manuscript collection. However, a significant part of NCDC's non-digital collection also exists at the FRC in East Point, Georgia. It is recommended that a detailed inventory of this collection be performed to create a comprehensive inventory of the entire manuscript collection.

These comprehensive inventories can then be used to accurately prioritize NCDC's manuscript records in regard to a rescue schedule. The rescue schedule can be divided into two related areas of vital work addressed below:

1) NCDC has maintained an aggressive micrographic preservation program since the 1950s whereby many paper records were operationally film as they were received. In addition, a special four year project began in the late 1970s and resulted in the creation of microfiche for many of NCDC's older records. Many of the records were destroyed after filming. However, some high priority records were retained due to uncertainty over maintaining a 100 percent accuracy level during the filming process.

A plan needs to be developed and then implemented which determines the uniqueness between NCDC's film and manuscript collections. For those records that exist on both film and paper, the plan should address the disposal of one of the duplicate medias. The records that were destroyed and exist on film only represent the rescue challenge described in section 4.2.2.

2) Once the question of duplication of film and paper records is answered, an aggressive preservation program needs to be established to rescue those unique manuscript records at greatest risk and of highest priority. The program should involve the scanning of the images as described in section 4.2.2. In 1995, NCDC began this process with the station history manuscript records which is a small but very high priority collection.

4.2.2 Micrographic Records

Accomplishments: A critical rescue situation exists with NCDC's micrographic or film collection. Few accomplishments have been made thus far in rescuing these data except for physically separating the contaminated film.

Based upon information and test results collected in 1994 by NCDC from the National Institute of Standards and Technology, the National Media Lab, and the Image Permanence Institute, NCDC's vast acetate film collection is in an advanced state of deterioration with less than three years of remaining useful life. The film is emitting acetic acid. The main physical symptoms are a strong vinegar odor and brittleness/ buckling of the film itself. The chemical reaction created by the acetic acid exponentially breeds upon itself, cannot be stopped once it starts, and infects "clean" film stored in the same physical area. The problem was first noticed in the early 1990's when a slight vinegar odor was detected. By 1994, the vinegar odor became so intense that a health hazard was declared and the storage rooms were sealed from human contact until forceful ventilation systems could be installed.

The degrading film contains most original record forms of the historical U.S. climate observation network prior to the 1970's. It was created operationally from the 1950's through the 1970's. In November 1977, NCDC converted the daily operational film production to a polyester based microfiche media which is not affected by this degradation. The historical data on 16mm and 35mm film was not migrated to the newer media at that time since it was still considered to be a stable media.

A small sample of the type of information available on this film includes all original forms of rawinsonde and pibal reports from the 1940's and 1950's, ship observations from the 1800s through the 1940s,

weather diaries from the 1700's and 1800s, barogram and thermograph charts from the 1800's through the 1950s, solar radiation forms from the 1940's and 1950s, as well as forms and summaries from various older research experiments such as IGY, BOMEX, FGGE, and GARP. Some of the original paper records have been destroyed making these film records the only available copy. In addition, much of the information contained on the film has never been keypunched.

Vital Work: This task will rescue approximately 100,000 rolls of 35mm and 16mm film from imminent danger of being lost due to acetate-based film degradation. The rescue process requires a multi-phased approach that will result in preserving the highest priority film first. Thus, the first year of the project will involve the development of a roll-level inventory in order to prioritize the film collection and eliminate duplicate holdings or film with little value. The subsequent years of the project will result in the scanning and indexing of each film image and preserving the resulting image on magnetic media. Thus, the project will involve the following steps:

1) separate the contaminated and uncontaminated film and store in separate rooms. Slow down the process of film degradation by improving environmental conditions (ventilation, lower temperature and humidities). (This was completed 2/1995)

2) perform an inventory of the 100K rolls of film to the resolution of a roll. Among other items, the inventory for each roll will include: a) Data Set (eg. thermograph chart, pibal, research experiment, etc.), b) Roll Content/ Description (eg. station name, experiment name, etc.) c) Begin and End Dates

3) using the inventory created in step 1), prioritize the film records based upon current research requirements, projected future needs, and availability of the records from other sources. Also during this stage, rolls of film which are of little or no value and rolls which are duplicate with other holdings will be recommended for disposal. Present the priority list to NCDC's Advisory Panel for their recommendations and approval.

4) begin the scanning and imaging process for the high priority film continuing the effort until the project is completed. The industry standard TIFF Group IV format will be employed for the conversion of the document images. It is estimated that approximately 1,000 images are contained on each roll of film. NCDC has received estimates of 10 cents per image to perform this process. OCR will not be performed due to the high cost ($1.00 per image) and also the lack of accuracy with current technology. For those high priority rolls that are already badly degraded, determine if a better copy exists at NCDC's off-site back-up storage facility for use in scanning.

5) transfer the scanned images to 3590 magnetic media. High priority images will be made accessible to users via NCDC's Hierarchical Data Storage System. This is a mass storage system which allows on-line access to NCDC's digital climate data via 3590 cartridges.

The project will result in the migration from film media to magnetic media. Plans are to use 3590 magnetic tape cartridges to hold the scanned digitized images. Each cartridge has a capacity of 10 gigabytes uncompressed and currently costs $50. per unit. This media is the successor to the IBM 3480/3490 cartridge family which is currently accepted by NARA as approved archive media.

It is estimated that a maximum of 10 terabytes of digital images will be created (100 kilobytes per image) if all 100,000 rolls of film are migrated. The actual number of terabytes rescued will be less and will be determined during the first year of the effort after the inventory/ prioritization phase of the project is completed. Optical media was considered but will not be used due to the lack of industry standards currently available.

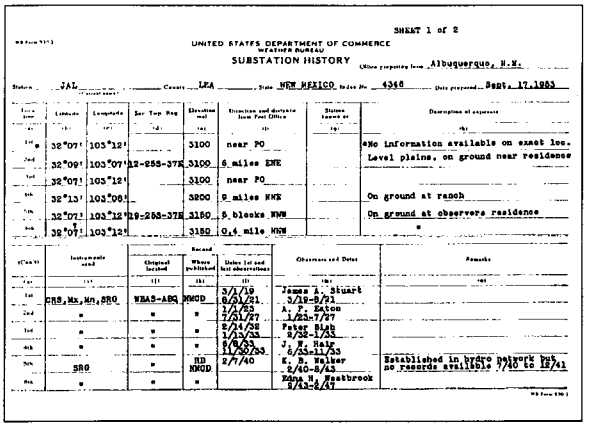
4.3 Metadata

4.3.1 Station History

Accomplishments: Access to histori-

cal environmental data is impeded by the lack of a single, consolidated, and accurate source of station history information. Volunteer observers have provided over a century's worth of meteorological data. However, much of these data remain untapped, primarily because no one readily knows of their existence.

As a result of ESDIM, access to station history information for approximately 20,000 NWS Cooperative Stations is being been made available. Continuous station history information for the 1890-1948 period, as recorded on the Weather Bureau Substation History forms (WB 530s), are currently being rescued. As shown in Figure 9, the WB 530 chronicles the changes in station locations over time for surface cooperative sites. This includes changes in latitude, longitude, elevation, and instrumentation. Prior to this project, station history information was available in digital form only since 1948 for these surface sites. This data base is now operationally maintained as part of NCDC's base funds.



SHEET 1 of 2

UNITED STATES DEPARTMENT OF COMMERCE
WEATHER BUREAU
SUBSTATION HISTORY

Station JAL County LEA State NEW MEXICO Index No. 4346 Date prepared Sept. 17,1985

Line	Latitude	Longitude	Sec. Twp. Rng	Elevation msl	Direction and distance from Post Office	Station moved to	Description of exposure
1st	32°07'	103°12'		3100	near PO		No information available on exact loc.
2nd	32°09'	103°07'	12-25S-37E	3100	6 miles ENE		Level plains, on ground near residence
3rd	32°07'	103°12'		3100	near PO		
4th	32°13'	103°06'		3200	9 miles NNE		On ground at ranch
5th	32°07'	103°12'	19-26S-37E	3150	8 blocks NNW		On ground at observers residence
6th	32°07'	103°12'		3150	0.4 mile NNW		

Line	Instruments used	Original furished	Where published	Dates 1st and last observations	Observers and Dates	Remarks
1st				3/1/10	James A. Stuart	
2nd	CRS,Mx,Mn,SRG	WBAS-ASO	NMOD	6/21/21	3/19-8/21	
3rd	"	"	"	1/1/23	A. F. Eaton	
4th	"	"	"	7/31/27	1/23-7/27	
5th	"	"	"	2/21/28	Peter Rich	
6th	"	"	"	1/15/33	8/28-1/33	
7th	"	"	"	6/9/33	J. W. Hale	
8th	"	"	"	11/30/33	6/33-11/33	
9th	SRG	"	NB NMOD	2/7/40	K. B. Walker 2/40-8/43	Established in prior network but no records available 7/40 to 12/41
10th	"	"	"		John W. Westbrook 8/43-8/47	

Figure 9: *An example of an old form in NCDC archives being rescued. This is a WB 530 station history form. It contains metadata about one observation site. About 20,000 forms are being digitized to extend NCDC's station history data base back to 1890. These metadata are critical to global change researchers in the elimination of biases in the climate record.*

Use of station history information provides researchers a means to assess the homogeneity of the meteorological data used to prepare baseline data sets. Station relocations (moves, changes in exposure, instrumentation, observers, etc.) can introduce discontinuities into the climatological data records. Researchers of global climate change depend upon this ancillary data to detect biases and remove inhomogeneities prior to the production of baseline climate data sets.

The rescue and access are being accomplished through five steps:

1) "Pre-key edit" the WB 530s--prepare the copies of the forms
2) Digitize the WB 530s utilizing Contractor Services
3) Post-edit the keyed information
4) Merge with post-1948 NCDC Station History Data Base
5) Develop an on-line inventory of these metadata

At the end of fiscal year 1995, about 90 percent of steps 1 through 3 will be completed. The final 10 percent and steps 4 and 5 remain to be accomplished. Unfortunately, ESDIM funding was not extended for an additional year to complete this work.

Vital Work: ESDIM funding needs to be extended for two additional years to complete the 1890-1948 data rescue for surface cooperative sites as described above. This is a critical station history milestone that should be completed. A number of additional tasks remain to be done and are outlined below:

1) A station history data base exists in manuscript form containing all stations that existed in the United States from the early 1800's through 1890. This book was published by the U.S. Signal Service in 1891. By using similar procedures as described above for the 1890-1948 period, the station history data base could be extended back in time to include all stations that were operating from the early 1800's through the current. This is a high priority project that should be completed in order to aid in current climate and global change research.

2) The development of a station history data base for the upper air network is critical for climate change research. Currently, the CARDS project at NCDC has made a first cut at developing a version 1 level data base. The data base needs to be expanded and enhanced.

3) It is also proposed to capture the times of observation for various meteorological elements and supplement wherever possible with metadata from other sources of station history. Observation time information is not digitally available prior to the 1980s. The most reliable sources for observation times are the manuscript records of meteorological data (WB 1009s) and the "Report on Substation" forms (WB B-44s).

4) There is a need to research, digitize, and merge historical first order station instrumentation into the master data base. Climate continuity issues depend on this type of metadata to determine possible biases caused by changes in instrumentation through time.

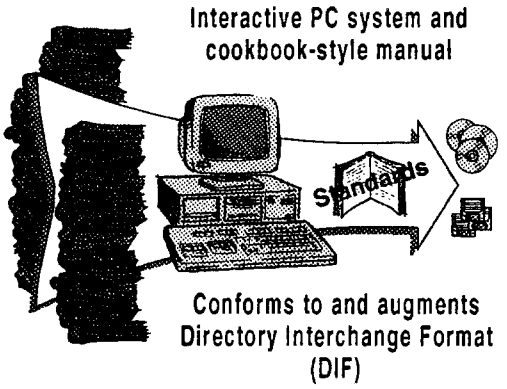


Figure 10: *NCDC's work in developing an expert system that builds a 10 to 20 page data set documentation manual. The expert system guides users through the documentation process. Much work remains to be done here.*

5) Comparisons need to be made between inventories of non-digital and digital data with the station history data base. NCDC currently maintains the station history and these inventories as separate files. The purpose of the comparison would be to identify time periods when the inventories show that data exist for a station but the station history indicates that station was not in operation. The station history data base would then be updated to reflect the years that data actually exist.

4.3.2 Data Set Documentation

Accomplishments: NCDC recently created a PC-based expert system under the Climate & Global Change program which can be used to create standardized documentation for all surface climate data sets. A graphic of this system is shown in Figure 10. The expert system is divided into 58 questions or topics. When these questions are answered a data set will be completely documented. These topics are totally compatible with the DIF used in the NOAAESDD and NASAGCD. In fact, one of the two outputs from the expert system is a DIF entry for the data set. Since the DIF entry is just a one-page overview of the data set, the second product is the complete documentation manual which is usually 10 to 20 pages in length. This latter product is critical for researchers interested in some of the finer details associated with a data set (e.g., instrument biases, quality control techniques, known data set problems, etc.).

The development of standard documentation practices is critical to allow users the ability to gain the information necessary to adequately understand a data set. This is very important to contemporary research involving the use and management of climate data. The idea behind the development of these standards is to produce documentation which passes the 20-year test. This test asks the question: Will anyone be able to understand or use a contemporary data set 20-years from now? The answer is usually no since data base developers usually do not know how to prepare useful documentation. In most cases, this is due to the lack of standards. This can result in disorganized

pieces of information, sometimes scattered in many places, and often existing only on single copy paper media. Standards make it easier for developers of data bases to produce pertinent documentation by asking a series of the right questions which then serve as a reference. This also results in a digital record of the documentation, thus rescuing single copy paper media from potential loss.

Vital Work: At present, NCDC's digital archives contain over 220 terabytes of information. Several hundred digital data sets are available on over 450,000 units of magnetic media. Many of these data sets are no longer updated and contain only historical information. Unfortunately, complete documentation and descriptive information have been prepared only for the larger and more contemporary data sets, and are usually incomplete or exist on single copy paper media for many of the older and but potentially important collections. As with most metadata, the documentation that does exist has been developed over many years and by many different organizations. This has resulted in heterogeneous metadata due to the lack of standard methods for creating it.

In order to rescue documentation from the risk of potential loss and also provide adequate descriptive information to data holdings important to contemporary research, a multi-phased plan is required:

1. Immediately rescue documentation which exists in single copy paper media by making a back-up copy.

A survey will be made of the existing data sets which exist only on single copy media. The documentation for these data sets will be photocopied. In addition, an off-site back-up storage area will be created where all back-up copies of documentation will be maintained. The off-site storage area will likely be in the same area as the NCDC off-site magnetic media library. This will allow easy access and also incur only a minimal expense for the documentation storage space.

2. Prepare official documentation and descriptive information on NCDC data holdings using the PC-based expert documentation system.

For those data sets which require documentation, the information available for each data set will be reviewed, missing information necessary for preparation will be identified, and contacts will be made and multiple sources researched to obtain it. Information will be presented in a concise, comprehensive manner using the PC-based expert documentation system.

5.0 THE COST: Five Year Data Rescue Budget

Data rescue is a dynamic facet of data management that requires continuous planning and a commitment of regular funding. Significant progress has been made in preserving important data at NCDC. A summary of these accomplishments was presented in Section 4 and is also shown in Figure 11. Due to the volume and diversity of NCDC data, much rescue work remains to be completed.

ACCOMPLISHMENTS

- ✔ Over 222,000 magnetic tapes migrated
- ✔ Off-site back-up of all migrated tapes
- ✔ 3% random sampling of digital data
- ✔ NEXRAD ingest system operational
- ✔ ASOS ingest system operational
- ✔ 100K FOSDIC images converted
- ✔ Box inventory of manuscript records
- ✔ Manuscript archival / retrieval system
- ✔ Inventory of all FDL publications
- ✔ Physical separation of contaminated film
- ✔ Over 3.5 million COADS records keyed
- ✔ 1890-1948 station history 90% complete
- ✔ Expert documentation system created

Figure 11: *Data rescue accomplishments at NCDC over the last 5 years*

A comprehensive listing of the vital work still to be done is listed in Table 2. The table depicts 26 separate rescue tasks along with the section in this report where the associated work is described. The tasks have been prioritized based upon the methodology described in Section 3.

Table 3 presents the costs associated with this vital work through Fiscal Year 2000. The tasks are presented from highest to lowest priority order. To accomplish these tasks, the incremental cost over and above NCDC base funding is $2.7 million for five years. It is important to note that by the year 2000, the operational ingest of NEXRAD will be $600K or approach one fourth of the entire rescue budget. Also, after the year 2000, the disposition of massive NEXRAD archive will need to addressed in terms of migrating all or part of the data to a newer media.

In order to conduct this proposed rescue program at NCDC, it is recommended that a rescue administrator position be added at NCDC to manage these tasks and associated contract work.

Table 2. Prioritized listing of 26 rescue tasks.

Priority	*Rescue Category / Subcategory*	*Section of Report*	*Task Description*
A	All	3.3	Comprehensive records schedule
A	All	3.3	3% random sampling
A	Digital / Magnetic	4.1.1	Round tape migration
A	Digital / Magnetic	4.1.1	3480 migration
C	Digital / Magnetic	4.1.1	GOES back-ups
A	Digital / New data	4.1.2	NEXRAD operations
A	Digital / New data	4.1.2	ASOS operations
B	Digital / New data	4.1.2	Lightning ingest
B	Digital / New data	4.1.2	GOES cloud cover
B	Digital / New data	4.1.2	ASOS supplemental observations
A	Digital / FOSDIC	4.1.3	Image conversion
A	Non digital / Manuscript	4.2.1	COADS keying
B	Non digital / Manuscript	4.2.1	FRC inventory
B	Non digital / Manuscript	4.2.1	Unique versus duplicate records
C	Non digital / Manuscript	4.2.1	Image scanning
A	Non digital / Film	4.2.2	Inventory
A	Non digital / Film	4.2.2	Prioritization
C	Non digital / Film	4.2.2	Image scanning
A	Metadata / Station History	4.3.1	WB 530 keying
B	Metadata / Station History	4.3.1	Pre-1890 keying
A	Metadata / Station History	4.3.1	Upper air development
C	Metadata / Station History	4.3.1	Observation time keying
C	Metadata / Station History	4.3.1	Instrumentation keying
C	Metadata / Station History	4.3.1	Inventory-station history compare
A	Metadata / Document	4.3.2	Single copy rescue
B	Metadata / Document	4.3.2	Documentation preparation

Table 3. Associated costs for 26 rescue tasks.
($ in thousands)

Priority	*Rescue Category / Subcategory — Task Description*	*FY96*	*FY97*	*FY98*	*FY99*	*FY00*
A	All — Comprehensive records control schedule	$50	$ -	$ -	$ -	$ -
A	All — 3% random sampling	50	60	65	75	90
A	Digital / Magnetic — Round tape migration	720	-	-	-	-
A	Digital / Magnetic — 3480 migration	80	800	810	850	850
A	Digital / New data — NEXRAD operations	300	400	500	550	600
A	Digital / New data — ASOS operations	250	250	120	130	145
A	Digital / FOSDIC — Image conversion	50	-	-	-	-
A	Non digital / Film — Inventory	75	-	-	-	-
A	Non digital / Film — Prioritization	85	-	-	-	-
A	Non digital / Manuscript — COADS keying	75	75	80	85	90
A	Metadata / Station History — WB 530 keying	110	-	-	-	-
A	Metadata / Station History — Upper air development	85	50	-	-	-
A	Metadata / Document — Single copy rescue	55	-	-	-	-
B	Digital / New data — Lightning ingest	115	77	50	25	25
B	Digital / New data — GOES cloud cover	50	30	-	-	-
B	Digital / New data — ASOS supplemental observations	100	25	-	-	-
B	Non digital / Manuscript — FRC inventory	-	-	100	100	-
B	Non digital / Manuscript — Unique versus duplicate records	-	75	75	-	-
B	Metadata / Station History — Pre-1890 keying	100	-	-	-	-
B	Metadata / Document — Documentation preparation	60	70	90	90	90
C	Digital / Magnetic — GOES back-up	-	200	225	240	250
C	Non digital / Manuscript — Image scanning	200	250	250	275	275
C	Non digital / Film — Image scanning	90	300	320	340	360
C	Metadata / Station History — Observation time keying	-	50	50	-	-
C	Metadata / Station History — Instrumentation keying	30	30	50	-	-
C	Metadata / Station History — Inventory-station history compare	50	-	-	-	-
	Totals:	**$2,750**	**$2,742**	**$2,785**	**$2,760**	**$2,775**

Appendix A

Procedure to Migrate Magnetic Media at Risk at the National Climatic Data Center

Migration Procedure Under Ideal Conditions

Assuming that the appropriate hardware and media type have been researched and then procured, the migration of digital reels which are in danger of being lost can generally be divided into three steps:

1. Solicit the advice of internal and/or external experts to determine the priority of data sets to be rescued. Special emphasis should be placed on those data which are important to ongoing and also future programs. In regard to NOAA, this includes data sets which are important to climate and global change research. Divide all data sets into two groups: those data sets which have media 5 years of age or older and those data sets which have media less than 5 years old.

2. Data sets which have media 5 years of age or older should be migrated first. These endangered data sets should be prioritized as described in step 1 above and migrated in the priority order. At NCDC, this includes data sets important to current and projected research such as CARDS, COADS, and past global weather experiments.

3. Data sets which have media less than 5 years old are migrated last. This group is also prioritized based upon importance. However, consideration must be given to the dynamic nature of many of these data sets. In other words, it is extremely important that the historical data as well as updates are migrated concurrently. If this is not done, processing difficulties will occur when using various software utilities such as sorts or selects.

All reels which are migrated should also have backups using the same media type. Also, the original media should be retained for an arbitrary period of time to guard against data loss which may have occurred during the migration or due to potential unforeseen degradation of the new media. A recommended retention period is five years. A staggered migration schedule should be devised for future migrations so that magnetic media does not become older than five years. Also, random sampling should be performed annually of all digital media. A 3 percent sample is recommended.

Magnetic media can be safely migrated using several methods. The following method is recommended if tapes are also consolidated:

1. Perform an initial tape evaluation (TVAL) which will identify the number of files, mode, format, and data volume.
2. Prepare a runstream which will consolidate as many input tapes onto a single output media as possible. Consolidation is based either upon data volume or some other prescribed plan (eg. 1 station or 1 month per tape). It is not recommended to split an input tape when migrating to the new output media.
3. Copy (ECOPY) the input reel(s) to the new output media.
4. Perform a character for character or bit for bit comparison of all files on the original reel(s) with new output media.
5. Create a backup of the new media and perform a bit for bit compare (between Library and Backup reels).

Migration Procedures Followed at NCDC

With some exceptions, NCDC followed the ideal plan described above. The priority of migrated data sets was devised mainly from in-house expertise although some external advice was solicited. Data sets were grouped into endangered and regularly accessed categories although we did deviate from the ideal migration plan due to the hardware and media limitations.

Tape cartridges were new to NCDC in 1990. At that time there was some question concerning degradation of

the metallic compounds used in the media. As a result, some of the lowest priority data sets in the endangered category were migrated first in order to provide a test of the media. These data sets included TD1440 (Surface Hourly), TD5600 (Upper Air), and TD9727 (Coop Summary of the Day). These were migrated during the Spring and Summer of 1990. All of these data sets are available in newer formats and are no longer accessed by users. The data sets are retained to guard against the potential loss of data due to processing problems which may have occurred during the reformatting in the 1980's.

In order to provide another test of the new media, our highest access data sets from the second category (less than 5 years old) were migrated during the Spring of 1991. These include the TD32 series such as Summary of the Day, Hourly Precipitation, and Surface Hourly. This was done in order to allow users to test the new media by frequent access. At the same time, we also migrated the high priority data sets in the endangered category. These include TD9939 (Service Record Retention Systems) and data sets from GATE, FGGE, STREX and ALPEX.

A major limitation which caused us to deviate from the ideal plan is and has been the lack of hardware (initially cartridge drives and now tape drive controllers). As stated above, both historical data and updates to data sets must be migrated concurrently in order to allow efficient user service. This has required us to migrate many of our static data bases first since we could not rely on hardware to be available to operationally migrate updates without impacting true data rescue.

The lack of hardware has also required us to revert back to allowing user access only to the original round reels to free-up competition for the available cartridge drives. Unfortunately, this defeats the purpose of migrating the TD32 series for testing the new media as stated above.

Appendix B

Random Sampling Procedures of Digital Data at the National Climatic Data Center

Introduction:

The National Archives and Records Administration (NARA) states that "Agencies shall annually read a statistical sample of all reels of magnetic computer tape containing permanent and unscheduled records to identify any loss of data and to discover and correct the causes of data loss" (NARA, 1990). The random sampling program has two separate phases. The National Institute of Standards and Technology (NIST) states that "The first phase has the purpose of detecting and identifying any parts of the ... collection where serious degradation has occurred." "The purpose of the second phase follow-up study is to evaluate the extent of any problems discovered in the first phase and to recommend or initiate appropriate corrective action" (NIST, 1988).

An annual 3 percent random sample has been performed on NCDC's digital data as prescribed by NARA and NIST since 1991. This effort has been supported by NOAA's Earth System Data and Information Management (ESDIM) program.

Sampling Procedures:

A systematic sampling procedure is used at NCDC as described by NIST (1988). Briefly, the procedure is delineated in the following steps:

Step 1: The sampling interval (S) is determined according the equation S = 1/f where f is the fraction of the population to be sampled (i.e., .03 or 3 percent). Thus S equals 33 which indicates that the sample will consist of every 33rd unit from the population.

Step 2: A random starting number, R, is chosen between 1 and S by consulting a random number table or generator. Thus the first unit selected in the sample will be media unit R and the rest of the sample will be chosen as every 33rd media unit in the population series after

R.

Step 3: Tape evaluation software is used on each media unit from the sample in order to determine data loss.

Data Loss Identification

The first phase of the random sample has the purpose of detecting and identifying any parts of the collection where degradation has occurred. NCDC uses a UNISYS utility called TVAL which determines read errors on the media units chosen for the random sample. NARA (1990) requires that tapes with more than 10 read errors be replaced. NCDC has adopted more rigorous procedures than the minimum required by NARA. NCDC replaces any media unit which has one or more read errors. Also, in the creation of archive media units, NCDC replaces any media unit with one or more write errors.

Corrective Action to Avoid Data Loss

The purpose of the second phase follow-up study of the random sample is to evaluate the extent of any problems discovered in the first phase and to recommend or initiate appropriate corrective action. At NCDC, this second phase involves determining some pattern or logical link among the defective media units through the sharing of some common characteristics. Examples of common characteristics would include media units stored in the same storage location or media units created 1) from the same manufacturer, 2) on the same date, 3) from the same tape drive, 4) from the same software.

The identification of a common problem from the random sample would lead to follow-up checks of other media units in the entire population. If the problem is confirmed as non-random, then NCDC makes an attempt to correct the defective media units from the entire population, not just the statistical sample. If the problem is confirmed to be random, then attempts are made to correct only the defective unit(s) found from the statistical sample. In almost all instances the corrective action is made from back-up copies of the media unit. In the rare instance when there is no back-up or the back-up is also found to be defective, then a search is made for other digital sources of the data at other institutes. In the event that no other digital data source is found and the data are deemed to be of high priority, the data may be re-digitized from manuscript sources if applicable.

References:

NARA, 1990: Managing Electronic Records. National Archives and Records Administration, Washington, DC.

NIST, 1988: Survey Sample Design for Microfilm Inspection at the National Archives. MISTIR 88-3889. National Institute of Standards and Technology. Gaithersburg, Maryland.

List of Acronyms

ASOS	Automated Surface Observation System
BOMEX	Barbados Oceanographic Meteorological Experiment
C&GC	Climate & Global Change Program
CZCS	Coastal Zone Color Scanner
CARDS	Comprehensive Aerological Reference Data Set
COADS	Comprehensive Ocean-Atmosphere Data Set
DIF	Directory Interchange Format
EOSDIS	Earth Observing System Data Information System
ESDIM	Environmental Services Data and Information Management
FDL	Foreign Data Library
FGGE	First GARP Global Experiment
FOSDIC	Film Optical Sensing Device for Input to Computer
FRC	Federal Records Center
GAO	General Accounting Office
GARP	Global Atmospheric Research Program
GOES	Geostationary Orbiting Environmental Satellite
GSA	General Services Administration
HDSS	Hierarchical Data Storage System
HRPT	High Resolution Picture Transmission
IG	Inspector General
IGY	International Geophysical Year
ISCCP	International Satellite Cloud Climatology Project
NARA	National Archives and Records Administration
NASAGCD	NASA Global Change Directory
NBS	National Bureau of Standards
NCDC	National Climatic Data Center
NEXRAD	Next Generation Radar
NOAA	National Oceanic and Atmospheric Administration
NOAAESDD	NOAA Earth Systems Data Directory
NWS	National Weather Service
OCLC	On-Line Catalog Library Center
OCR	Optical Character Recognition
POES	Polar Orbiting Environmental Satellite
SSM/I	Special Sounder Microwave/Infrared
WB 530	Weather Bureau Form 530 Substation History
WSR-88D	Weather Surveillance Radar 1988 Doppler